instituto
de ingeniería
del conocimiento

Benchmarking LLMs on Custom Spanish Corpus

Author:

David Betancur
Data Scientist
Instituto de Ingeniería del Conocimiento

instituto de ingeniería del conocimiento

Benchmarking LLMs on Custom Spanish Corpus

In our search for the technology that best suits our client's needs, we recently conducted a comprehensive **benchmark to evaluate several large language models** (LLMs) for Generative AI across different linguistic corpora in Spanish. We decided to focus on models in the 7–12 billion parameter range as they offer an ideal balance between language capabilities and cost-effectiveness.

As a consequence, these are some of the most popular models among developers and enterprise customers. In particular, they can run on AWS's G5 instances (which are cheaper than the high-end instances required for larger models) while still delivering competitive performance. Among these, StabilityAI's Stable LM 2 model (`stabilityai/stablelm-2–12b-chat`) stands out by consistently delivering superior results over similarly sized models.

David Betancur
Data Scientist
Instituto de Ingeniería del Conocimiento

Thanks to:
Álvaro Barbero Jiménez
Pablo Haya
Marta Guerrero Nieto
Kateryna Sushkova
Carmen Muñoz
Natàlia López
Nuria Aldama
Helena Montoro
for the support and development of the corpus used in this benchmark.

Also thanks to
Carlos Riquelme
for providing support to this project

1. INTRODUCTION

As part of our mission to implement cutting-edge solutions, we performed a rigorous benchmarking process involving multiple language models. This analysis spanned various datasets with the goal of identifying the model that delivers the most semantically accurate responses in Spanish in relevant and practical scenarios. The models compared include those from leading AI organizations such as OpenAI, Meta, MistralAI, Microsoft, and Stability AI.

2. METHODOLOGY

2.1. Spanish Annotated Q&A Corpora to Compare LLM

The evaluation was based on **four Spanish annotated linguistic corpora developed by the IIC**, covering a variety of scenarios from topical Q&A tasks to real-world business conversations. The corpora consist of manually created conversational question-answer pairs, as well as some that are semi-automatically generated and subsequently curated by IIC's team of computational linguists.

Specifically, we used three question-answering (QA) corpora and a fourth called 'citizen_information', which contains a series of questions in which spelling and writing errors have been gradually inserted. This degradation of the questions allows us to assess the sensitivity of the model to the variability of the user's actual questioning style.

The datasets used are as follows:

- **IIC/AQuAS:** Contains questions, answers, and contexts for a variety of domains. This dataset is publicly available at [1].
- **IIC/Retail:** Multi-turn conversations with support context for each question framed in the retail domain.
- **IIC/Insurance:** Multi-turn conversations with support context for each question focused on the insurance domain.
- **IIC/Citizen:** QA dataset from FAQ questions containing question degradations.

Our benchmarking corpora are designed to mimic Retrieval-Augmented Generation (RAG) tasks, and they evaluate the performance of LLMs in complex environments. In particular, these benchmarks help select the best model for in-domain use cases, such as supporting internal documentation inquiries in retail stores or aiding call center operators in responding to specific customer queries. Additionally, RAG systems are useful for controlling the model's tendency to hallucinate, ensuring more accurate and reliable responses.

2.2. Models

Each model was evaluated based on its ability to accurately generate responses. To measure the correctness of responses, we employed a Semantic Answer Similarity (SAS) metric [2] over the answers with respect to a reference created by human annotators. This metric employs another language model to assess the semantic similarity between the answer and the reference.

2.3. Hardware

Experiments were conducted on a p4 instance in AWS [3] to fit the largest model (Stable LM 2 12B) since the G5 instances were slightly smaller than the unquantized model.

3. RESULTS

Figure 1 shows the results of the QA benchmark where tasks are sorted in decreasing size (number of test examples). We see that Stable LM 2 12B outperforms the other models in every task on the benchmark.

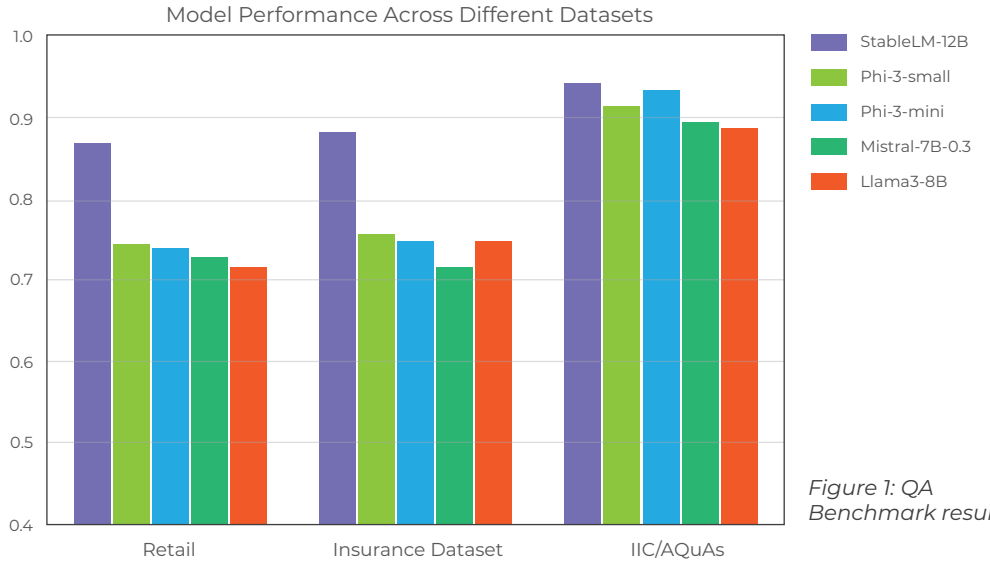
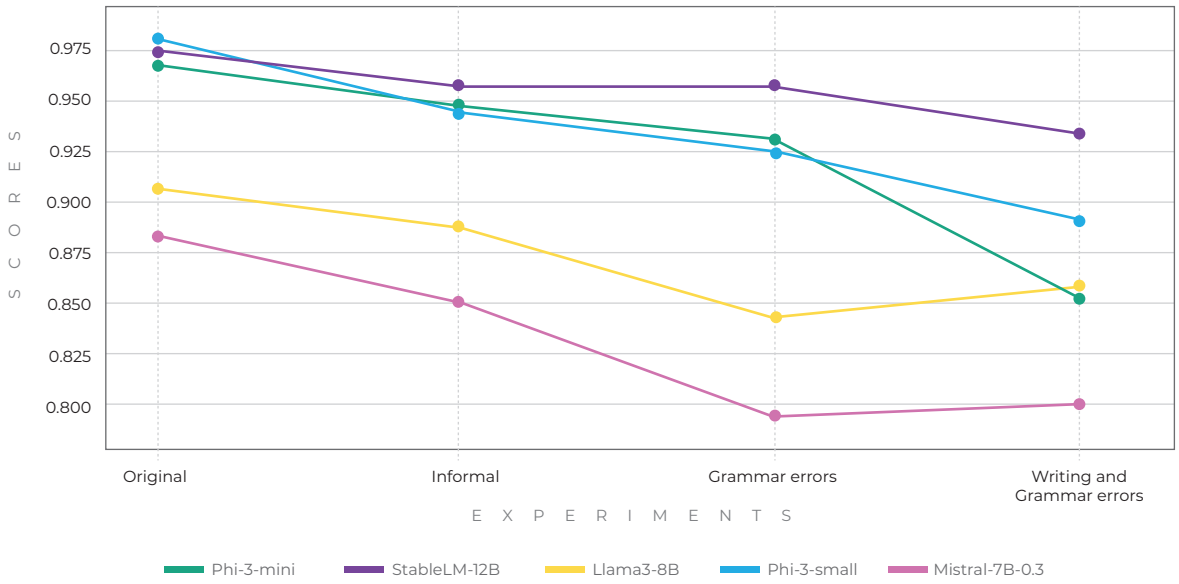


Figure 1: QA Benchmark results

- IIC/Retail Dataset (1326 questions and answers):
  - Stable LM 2 12B: 0.86
  - Other models, including Phi and Llama-3, performed notably lower at 0.74 and 0.73 respectively.
- IIC/Insurance Dataset (821 questions and answers):
  - Stable LM 2 12B: 0.88
  - The next best, Phi-Small, scored 0.75, demonstrating a clear gap in performance.
- IIC/AQuAS Dataset (107 questions and answers):
  - Stable LM 2 12B: 0.94
  - The closest competitor, Phi-3, scored 0.93, while others like Mistral and Llama3 lag further behind.

Figure 2 shows the results for the QA question degradation benchmark. A clear decrease in the answer quality is shown for every model, but some, such as Phi3-mini, have a steeper slope towards the end, showing sensitivity to question quality. On the other hand, Stable LM 2 12B maintains a low slope, meaning low sensitivity to questions with spelling or typing errors.

Figure 2: QA question degradation benchmark results



4. CONCLUSION

We conducted an in-depth benchmarking study centered around real-world applications in Spanish (retail, insurance, and citizens FAQs domains) using top language models in the 7–12 billion parameter range. Our results suggest that Stable LM 2 12B is the most promising model, followed by Phi3. Model evaluation is a challenging technical aspect of modern artificial intelligence, and we hope to continue making progress by linking benchmarks to practical use-cases in industry, especially in the sometimes overlooked multilingual setups.

5. REFERENCES

[1] IIC, IIC/AQuAS Dataset (2024), https://huggingface.co/datasets/IIC/AQuAS
[2] X. Zhang, Semantic Answer Similarity Metric (2021), https://arxiv.org/abs/2108.06130
[3] Amazon Web Services, P4 Instances (2024),

C/ Francisco Tomás y Valiente, nº 11
EPS, edificio B, 5ª planta
UAM Cantoblanco
28049 Madrid, España.

Tel.: (+34) 91 497 2323
Fax: (+34) 91 497 2334
iic@iic.uam.es
www.iic.uam.es